REMARKS

The present Amendment and the following Remarks are submitted in response to the Final Office Action mailed January 29, 2007. In this Amendment and Response After Final Rejection, Applicant is amending claim 30, canceling claim 59, and adding new claims 65 and 66. Claim 30 is being amended for consistency among the claim logic. Support for further amendment to claim 30 c) can be found in the specification at, for example, page 32, lines 15 to 18 (as amended on April 11, 2006). Support for the new claim 65 can be found in the specification at, for example, page 31, lines 2 to 3. This Amendment adds no new matter.

The minor amendments presented herein would not raise any new issues that would require further consideration and/or search. Applicant submits that the amendments would place the claims in condition for allowance or at least present the rejected claims in better form for consideration on appeal, and therefore shoud be entered after the final rejection under 37 C.F.R. §1.116. Claims 30, 31, 34, 36, 38, 56 and 60-66 will be pending upon entry of this amendment.

Applicant thanks the Examiner for withdrawing the objections and some rejections. The remaining points are addressed below.

Interview Summary

Applicant thanks the Examiner for the helpful telephone interview on March 29, 2007. The enablement rejection was discussed. An agreement on the approach to responding to the rejection was reached.

Paragraph 2. Restriction Requirement

The Examiner denied Applicant's request to examine claim 59 and maintained his assertion that it is properly included in a Group X outside of the group represented by the examined claims. In response, herein cancels claim 59 as being drawn to a non-elected invention. Applicant hereby reserves the right file a continuing application or take such other appropriate action as deemed necessary to pursue the subject matter of canceled claim 59 and/or claims in other groups. Applicant does not hereby abandon or waive any rights in the non-elected inventions.

Paragraph 8. Rejection of the Claims Under 35 U.S.C. §112, First Paragraph

Claims 30 and 56 remain rejected under 35 U.S.C. §112, first paragraph as allegedly containing subject matter which was not described in the specification in such a way as to enable one skilled in the art to which it pertains or with which it is most nearly connected, to make and use the invention commensurate in scope with the claims. In particular, the Examiner maintained that there would be undue experimentation for one of skill in the art to make or use an NCE1 protein at least 95% identical to SEQ ID NO:4 or encoded by a nucleic acid at least 95% identical to SEQ ID NO:3. Applicant respectfully traverses the rejection.

The Examiner asserted that there would be an undue amount of trial and error experimentation for one of skill in the art to search and screen a vast number of biological sources for any protein having at least 95% identity to SEQID NO:4 or encoded by any expression element comprising a nucleic acid sequence at least 95% identical to SEQ ID NO:3, wherein the protein forms a thioester linkage with NEDD8. Alternatively, the Examiner thought there would be too much trial and error experimentation for one of skill in the art to screen for specific residues or bases to change so the resulting protein does not have inactivated thioester linkage capabilities. Applicant respectfully disagrees.

First, a look at the scope of the claims. The NCE1 protein of SEQ ID NO:4 is 183 amino acids. To make a polypeptide at least 95% identical to the NCE1 protein (as in claim 30.a)), one could change up to about 9-10 amino acids. To make a polypeptide at least 99% identical to the NCE1 protein (as in new claim 65), one could change up to about 2 amino acids. To make a polypeptide encoded by a nucleic acid at least 99% identical to SEQ ID NO:3 (as in amended claim 30 c)), one could change up to about 6 bases, but would not necessarily be changing as many amino acid residues, due to the redundancy of different codons' abilities to encode the same amino acid. This is not an endless number of possibilities.

Furthermore, Applicant disagrees with the assertion that there will be trial and error experimentation. The specification, at page 31, lines 8 and 9, identified amino acid residue 111 of SEQ ID NO:4 as a residue important for retention of biological activity. The teachings of rational mutant design, such as in the previously cited WO95/18974, known to those skilled in the art include the use of alignments such as provided by Applicant at Figures 3 and 7. One of skill in the art would see the numerous conserved residues (80 identical residues in Figure 3 (aligned with one related enzyme) and 22 identical residues in Figure 7 (aligned with 16 related enzymes)). One of skill in the art would recognize that at least the 22 residues, if not the 80 residues, are more likely to be important for activity than others and thus preferably would change other residues to avoid destruction of activity.

In addition, Figures 3 and 7 could guide the skilled practitioner in making conservative amino acid changes as taught in WO95/18974. A comparison of the teaching at page 33, lines 3 to 22 of that publication with Figures 3 and 7 alignments show conservative differences of the NCE1 sequence with the other sequences. This would provide the skilled practitioner with two sets of information: 1) they would help identify conservative changes which may be tolerated at positions which align with other active enzymes and 2) the near conservation at similar residues could help identify residues where change is tolerated, but too much change might alter the activity to some degree and thus could steer the practitioner away from making too many changes of nearly conserved residues. When one views the alignments from the perspective of similarity in addition to identity, the number of trials is further decreased.

Finally, Applicant further submits, as Exhibit A., Bowie et al. ((1990) *Science* **247**:1306-1310), wherein the authors concluded that "proteins are surprisingly tolerant of amino acid substitutions (page 1306, col.2, lines 12-13). In these studies, authors performed approximately 1500 single amino acid substitution at 142 positions of the *lac* repressor and found that "about one-half of all the substitutions were phenotypically silent." Therefore, one can expect that a significant percentage of random substitutions in a given protein will result in mutated proteins with full or nearly full activity. These are far better odds than those at issue in *In re Wands,* 858 F.2d 731 (Fed. Cir. 1988), in which the court found that screening many hybridomas to find the few that fell within the claims was not undue experimentation. Based on Exhibit A's disclosure, one would predict that even random substitution of amino acid residues in NCE1 will result in a large pool of mutants having full or partial thioester linkage activity. When one makes less than random substitutions following the guidance of the Applicant in the specification and the teachings in the art including those cited in WO95/18974 (avoiding changes to evolutionally conserved amino acid residues and performing conservative substitutions), the odds of retaining function are greatly increased. Therefore, it is not undue experimentation to devise a polypeptide within the structural limitations of the claims having the functional limitation also found in the claims.

In summary, as discussed in the prior response, filed on November 7, 2006 (herein incorporated by reference) and supplemented here, the knowledge of one skilled in the art, supplemented by the structural information and functional assay methods provided by Applicants enables the production and identification of polypeptides encompassed by the claims (as amended) without undue experimentation. In view of the amendments and these remarks, Applicant respectfully requests that the rejection of claim 30 (claims 31 and 56 dependent thereon) be withdrawn.

Rejoinder

Applicant submits that claims 30, 31 and 56 (and new claims 65 and 66) will be found allowable, and the application is allowable with respect to the group elected after the Restriction Requirement mailed January 11, 2006. Applicant believes that now, the Examiner, under MPEP § 821.04, as noted in the Restriction Requirement, can undertake the review of the withdrawn process claims, 34, 36, 38 and 60-64, which depend from or otherwise include all the limitations of the allowable product claims. In the next Office communication, Applicant respectfully requests comment on these withdrawn claims after rejoinder.

CONCLUSION

The foregoing amendments and remarks are being made to place the Application in condition for allowance. Applicant respectfully requests that the Examiner consider these remarks after final rejection and indicate the allowance of the claims 30, 31 and 56 (and new claims 65 and 66) and rejoinder (and subsequent allowance) of claims 34, 36, 38 and 60-64 because, in view of these remarks, Applicants respectfully submit that the rejection of claim 30 (claims 31 and 56 dependent thereon) under 35 U.S.C. § 112 is herein overcome. Early notice to this effect is solicited.

If, in the opinion of the Examiner, a telephone conference would expedite the allowance of the subject application, the Examiner is encouraged to call the undersigned. If the Examiner disapproves of Applicants' amendments and/or remarks in this response, Applicants request a prompt mailing of an Advisory Action to that effect.

This paper is being filed timely within two months of the mailing date of the final action. No extensions of time are required. In the event any extensions of time are necessary, the undersigned hereby authorizes the requisite fees to be charged to Deposit Account No. 501668.

Entry of the remarks made herein is respectfully requested.

29 March 2007

Respectfully submitted,
MILLENNIUM PHARMACEUTICALS, INC.

By Tracy M. Sioussat
Tracy M. Sioussat, Ph.D.
Registration No. 50,609
40 Landsdowne Street
Cambridge, MA 02139
Telephone - 617-374-7679
Facsimile - 617-551-8820

Exhibit A to accompany
Amendment and Response After Final
in 10/681,690

Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions

JAMES U. BOWIE,* JOHN F. REIDHAAR-OLSON, WENDELL A. LIM,
ROBERT T. SAUER

**n amino acid sequence encodes a message that deter-
ines the shape and function of a protein. This message is
ghly degenerate in that many different sequences can
›de for proteins with essentially the same structure and
tivity. Comparison of different sequences with similar
essages can reveal key features of the code and improve
ıderstanding of how a protein folds and how it per-
rms its function.**

THE GENOME IS MANIFEST LARGELY IN THE SET OF PROteins that it encodes. It is the ability of these proteins to fold into unique three-dimensional structures that allows them to
ıction and carry out the instructions of the genome. Thus,
nprehending the rules that relate amino acid sequence to struc-
e is fundamental to an understanding of biological processes.
cause an amino acid sequence contains all of the information
:essary to determine the structure of a protein (1), it should be
ssible to predict structure from sequence, and subsequently to
er detailed aspects of function from the structure. However, both
›blems are extremely complex, and it seems unlikely that either
l be solved in an exact manner in the near future. It may be
ssible to obtain approximate solutions by using experimental data
simplify the problem. In this article, we describe how an analysis
allowed amino acid substitutions in proteins can be used to
uce the complexity of sequences and reveal important aspects of
ıcture and function.

ethods for Studying Tolerance to
quence Variation

:here are two main approaches to studying the tolerance of an
ino acid sequence to change. The first method relies on the
cess of evolution, in which mutations are either accepted or
cted by natural selection. This method has been extremely
verful for proteins such as the globins or cytochromes, for which
uences from many different species are known (2–7). The second
›roach uses genetic methods to introduce amino acid changes at specific positions in a cloned gene and uses selections or screens to identify functional sequences. This approach has been used to great advantage for proteins that can be expressed in bacteria or yeast, where the appropriate genetic manipulations are possible (3, 8–11). The end results of both methods are lists of active sequences that can be compared and analyzed to identify sequence features that are essential for folding or function. If a particular property of a side chain, such as charge or size, is important at a given position, only side chains that have the required property will be allowed. Conversely, if the chemical identity of the side chain is unimportant, then many different substitutions will be permitted.

Studies in which these methods were used have revealed that proteins are surprisingly tolerant of amino acid substitutions (2–4, 11). For example, in studying the effects of approximately 1500 single amino acid substitutions at 142 positions in *lac* repressor, Miller and co-workers found that about one-half of all substitutions were phenotypically silent (11). At some positions, many different, nonconservative substitutions were allowed. Such residue positions play little or no role in structure and function. At other positions, no substitutions or only conservative substitutions were allowed. These residues are the most important for *lac* repressor activity.

What roles do invariant and conserved side chains play in proteins? Residues that are directly involved in protein functions such as binding or catalysis will certainly be among the most conserved. For example, replacing the Asp in the catalytic triad of trypsin with Asn results in a 10^4-fold reduction in activity (12). A similar loss of activity occurs in λ repressor when a DNA binding residue is changed from Asn to Asp (13). To carry out their function, however, these catalytic residues and binding residues must be precisely oriented in three dimensions. Consequently, mutations in residues that are required for structure formation or stability can also have dramatic effects on activity (10, 14–16). Hence, many of the residues that are conserved in sets of related sequences play structural roles.

Substitutions at Surface and Buried Positions

In their initial comparisons of the globin sequences, Perutz and co-workers found that most buried residues require nonpolar side chains, whereas few features of surface side chains are generally conserved (6). Similar results have been seen for a number of protein families (2, 4, 5, 7, 17, 18). An example of the sequence tolerance at surface versus buried sites can be seen in Fig. 1, which shows the allowed substitutions in λ repressor at residue positions that are near the dimer interface but distant from the DNA binding surface of the protein (9). These substitutions were identified by a functional

authors are in the Department of Biology, Massachusetts Institute of Technology,
bridge, MA 02139.

ent address: Department of Chemistry and Biochemistry and the Molecular
gy Institute, University of California, Los Angeles, Los Angeles, CA 90024.

Fig
sho
que
low
wei
thr
and
frac
typ
to t
tide

sele
solv
sho
rela
posi
exp
chai
scen
prot

Co

B
for p
dict
only
prot
the l
show
tion
the l
sequ
Leu,
each
effec
resid
core
hydr
diffic
donc
possi
ly lin
the c
bond
Th
volur
volur
indiv
as sho
positi
chang

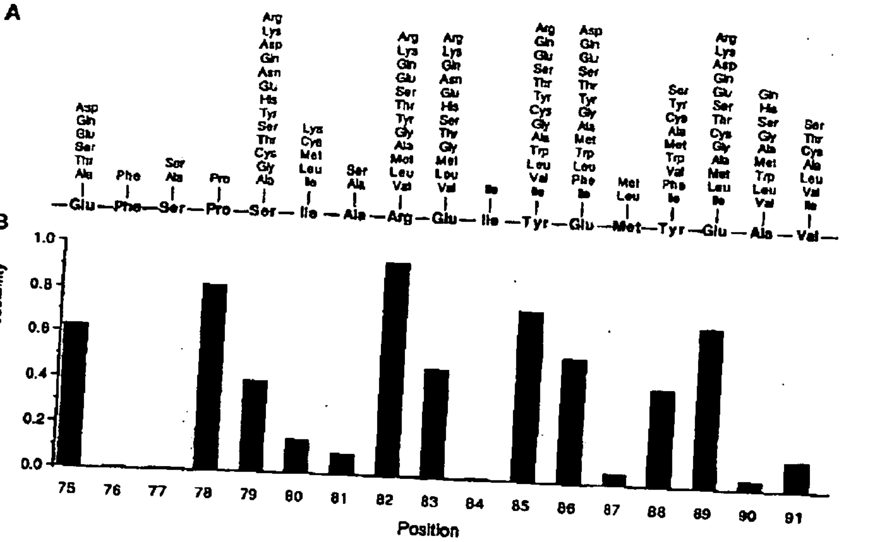

Fig. 1. (**A**) Amino acid substitutions allowed in a short region of λ repressor. The wild-type sequence is shown along the center line. The allowed substitutions shown above each position were identified by randomly mutating one to three codons at a time by using a cassette method and applying a functional selection (*9*). (**B**) The fractional solvent accessibility (*42*) of the wild-type side chain in the protein dimer (*43*) relative to the same atoms in an Ala-X-Ala model tripeptide.

selection after cassette mutagenesis. A histogram of side chain solvent accessibility in the crystal structure of the dimer is also shown in Fig. 1. At six positions, only the wild-type residue or relatively conservative substitutions are allowed. Five of these positions are buried in the protein. In contrast, most of the highly exposed positions tolerate a wide range of chemically different side chains, including hydrophilic and hydrophobic residues. Hence, it seems that most of the structural information in this region of the protein is carried by the residues that are solvent inaccessible.

Constraints on Core Sequences

Because core residue positions appear to be extremely important for protein folding or stability, we must understand the factors that dictate whether a given core sequence will be acceptable. In general, only hydrophobic or neutral residues are tolerated at buried sites in proteins, undoubtedly because of the large favorable contribution of the hydrophobic effect to protein stability (*19*). For example, Fig. 2 shows the results of genetic studies used to investigate the substitutions allowed at residue positions that form the hydrophobic core of the NH_2-terminal domain of λ repressor (*20*). The acceptable core sequences are composed almost exclusively of Ala, Cys, Thr, Val, Ile, Leu, Met, and Phe. The acceptability of many different residues at each core position presumably reflects the fact that the hydrophobic effect, unlike hydrogen bonding, does not depend on specific residue pairings. Although it is possible to imagine a hypothetical core structure that is stabilized exclusively by residues forming hydrogen bonds and salt bridges, such a core would probably be difficult to construct because hydrogen bonds require pairing of donors and acceptors in an exact geometry. Thus the repertoire of possible structures that use a polar core would probably be extremely limited (*21*). Polar and charged residues are occasionally found in the cores of proteins, but only at positions where their hydrogen bonding needs can be satisfied (*22*).

The cores of most proteins are quite closely packed (*23*), but some volume changes are acceptable. In λ repressor, the overall core volume of acceptable sequences can vary by about 10%. Changes at individual sites, however, can be considerably larger. For example, as shown in Fig. 2, both Phe and Ala are allowed at the same core position in the appropriate sequence contexts. Large volume changes at individual buried sites have also been observed in phylogenetic studies, where it has been noted that the size decreases and increases at interacting residues are not necessarily related in a simple complementary fashion (*5, 7, 17*). Rather, local volume changes are accommodated by conformational changes in nearby side chains and by a variety of backbone movements.

The Informational Importance of the Core

With occasional exceptions, the core must remain hydrophobic and maintain a reasonable packing density. However, since the core is composed of side chains that can assume only a limited number of conformations (*24*), efficient packing must be maintained without steric clashes. How important are hydrophobicity, volume, and steric complementarity in determining whether a given sequence can form an acceptable core? Each factor is essential in a physical sense, as a stable core is probably unable to tolerate unsatisfied hydrogen bonding groups, large holes, or steric overlaps (*25*). However, in an informational sense, these factors are not equivalent. For example, in experiments in which three core residues of λ repressor were mutated simultaneously, volume was a relatively unimportant informational constraint because three-quarters of all possible combinations of the 20 naturally occurring amino acids had volumes within the range tolerated in the core, and yet most of these sequences were unacceptable (*20*). In contrast, of the sequences that contained only

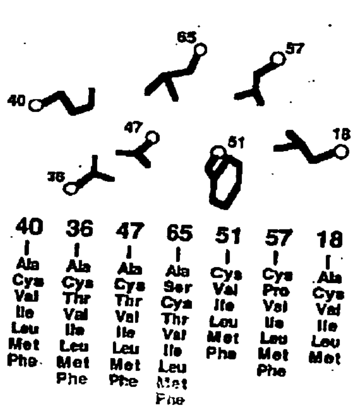
65
57
40
47
51
18
36
40 Ala Cys Val Ile Leu Met Phe
36 Ala Cys Thr Val Ile Leu Met Phe
47 Ala Cys Thr Val Ile Leu Met Phe
65 Ala Ser Cys Thr Val Ile Leu Met Phe
51 Cys Val Ile Leu Met Phe
57 Cys Pro Val Ile Leu Met Phe
18 Ala Cys Val Ile Leu Met

Fig. 2. Amino acid substitutions allowed in the core of λ repressor. The wild-type side chains are shown pictorially in the approximate orientation seen in the crystal structure (*43*). The lists of allowed substitutions at each position are shown below the wild-type side chains. These substitutions were identified by randomly mutating one to four residues at a time by using a cassette method and applying a functional selection (*20*). Not all substitutions are allowed in every sequence background.

the appropriate hydrophobic residues, a significant fraction were acceptable. Hence, the hydrophobicity of a sequence contains more information about its potential acceptability in the core than does the total side chain volume. Steric compatibility was intermediate between volume and hydrophobicity in informational importance.

The Informational Importance of Surface Sites

We have noted that many surface sites can tolerate a wide variety of side chains, including hydrophilic and hydrophobic residues. This result might be taken to indicate that surface positions contain little structural information. However, Bashford *et al.*, in an extensive analysis of globin sequences (*4*), found a strong bias against large hydrophobic residues at many surface positions. At one level, this may reflect constraints imposed by protein solubility, because large patches of hydrophobic surface residues would presumably lead to aggregation. At a more fundamental level, protein folding requires a partitioning between surface and buried positions. Consequently, to achieve a unique native state without significant competition from other conformations, it may be important that some sites have a decided preference for exterior rather than interior positions. As a result, many surface sites can accept hydrophobic residues individually, but the surface as a whole can probably tolerate only a moderate number of hydrophobic side chains.

Identification of Residue Roles from Sets of Sequences

Often, a protein of interest is a member of a family of related sequences. What can we infer from the pattern of allowed substitutions at positions in sets of aligned sequences generated by genetic or phylogenetic methods? Residue positions that can accept a number of different side chains, including charged and highly polar residues, are almost certain to be on the protein surface. Residue positions that remain hydrophobic, whether variable or not, are likely to be buried within the structure. In Fig. 3, those residue positions in λ repressor that can accept hydrophilic side chains are shown in orange and those that cannot accept hydrophilic side chains are shown in green. The obligate hydrophobic positions define the core of the structure, whereas positions that can accept hydrophilic side chains define the surface.

Functionally important residues should be conserved in sets of ictive sequences, but it is not possible to decide whether a side chain s functionally or structurally important just because it is invariant or onserved. To make this distinction requires an independent assay of rotein folding. The ability of a mutant protein to maintain a stably olded structure can often be measured by biophysical techniques, y susceptibility to intracellular proteolysis (*26*), or by binding to ntibodies specific for the native structure (*27, 28*). In the latter ises, it is possible to screen proteins in mutated clones for the ility to fold even if these proteins are inactive. Sets of sequences iat allow formation of a stable structure can then be compared to e sets that allow both folding and function, with the active site or nding residues being those that are variable in the set of stable oteins but invariant in the set of functional proteins. The DNA-nding residues of Arc repressor were identified by this method (*8*). e receptor-binding residues of human growth hormone were also ntified by comparing the stabilities and activities of a set of itant sequences (*28*). However, in this case, the mutants were erated as hybrid sequences between growth hormone and related mones with different binding specificities.

Implications for Structure Prediction

At present, the only reliable method for predicting a low-resolution tertiary structure of a new protein is by identifying sequence similarity to a protein whose structure is already known (*29, 30*). However, it is often difficult to align sequences as the level of sequence similarity decreases, and it is sometimes impossible to detect statistically significant sequence similarity between distantly related proteins. Because the number of known sequences is far greater than the number of known structures, it would be advantageous to increase the reach of the available structural information by improving methods for detecting distant sequence relations and for subsequently aligning these sequences based on structural principles. In a normal homology search, the sequence database is scanned with a single test sequence, and every residue must be weighted equally. However, some residues are more important than others and should be weighted accordingly. Moreover, certain regions of the protein are more likely to contain gaps than others. Both kinds of information can be obtained from sequence sets, and several techniques have

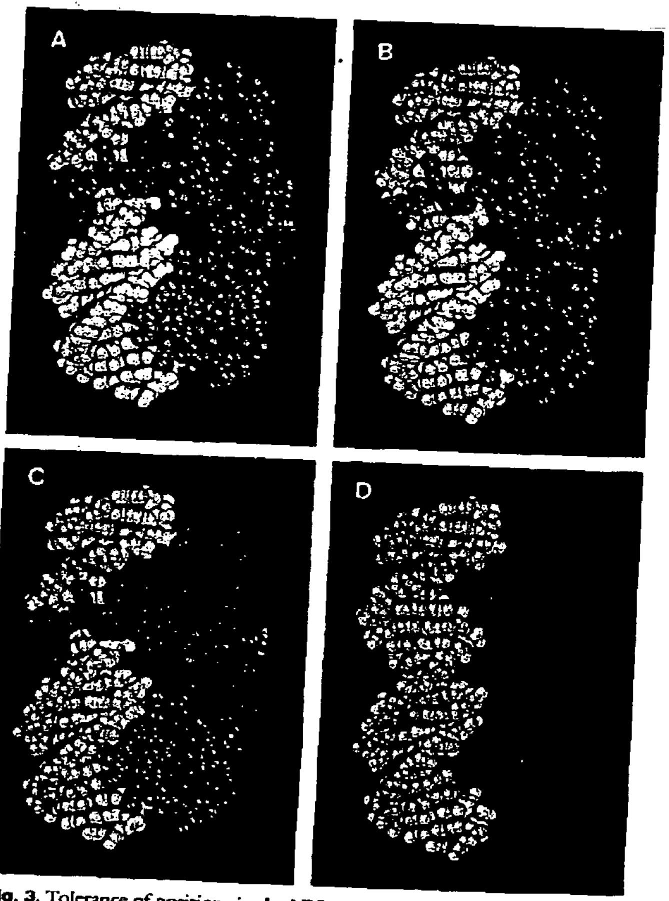

Fig. 3. Tolerance of positions in the NH_2-terminal domain of λ repressor to hydrophilic side chains. The complex (*43*) of the repressor dimer (blue) and operator DNA (white) is shown. In (**A**), positions that can tolerate hydrophilic side chains are shown in orange. The same side chains are shown in (**B**) without the remaining protein atoms. In (**C**), positions that require hydrophobic or neutral side chains are shown in green. These side chains are shown in (**D**) without the remaining protein atoms. About three-fourths of the 92 side chains in the NH_2-terminal domain are included in both (B) and (D). The remaining positions have not been tested. Data are from (*9, 14, 20, 27, 44*).

been used to combine such information into more appropriately weighted sequence searches and alignments (*31*). These methods were used to align the sequences of retroviral proteases with aspartic proteases, which in turn allowed construction of a three-dimensional model for the protease of human immunodeficiency virus type 1 (*29*). Comparison with the recently determined crystal structure of this protein revealed reasonable agreement in many areas of the predicted structure (*32*).

The structural information at most surface sites is highly degenerate. Except for functionally important residues, exterior positions seem to be important chiefly in maintaining a reasonably polar surface. The information contained in buried residues is also degenerate, the main requirement being that these residues remain hydrophobic. Thus, at its most basic level, the key structural message in an amino acid sequence may reside in its specific pattern of hydrophobic and hydrophilic residues. This is meant in an informational sense. Clearly, the precise structure and stability of a protein depends on a large number of detailed interactions. It is possible, however, that structural prediction at a more primitive level can be accomplished by concentrating on the most basic informational aspects of an amino acid sequence. For example, amphipathic patterns can be extracted from aligned sets of sequences and used, in some cases, to identify secondary structures.

If a region of secondary structure is packed against the hydrophobic core, a pattern of hydrophobic residues reflecting the periodicity of the secondary structure is expected (*33*, *34*). These patterns can be obscured in individual sequences by hydrophobic residues on the protein surface. It is rare, however, for a surface position to remain hydrophobic over the course of evolution. Consequently, the amphipathic patterns expected for simple secondary structures can be much clearer in a set of related sequences (*6*). This principle is illustrated in Fig. 4, which shows helical hydrophobic moment plots for the Antennapedia homeodomain sequence (Fig. 4A) and for a composite sequence derived from a set of homologous homeodomain proteins (Fig. 4B) (*35*). The hydrophobic moment is a simple measure of the degree of amphipathic character of a sequence in a given secondary structure (*34*). The amphipathic character of the three α-helical regions in the Antennapedia protein (*36*) is clearly revealed only by the analysis of the combined set of homeodomain sequences. The secondary structure of Arc repressor, a small DNA-binding protein, was recently predicted by a similar method (*8*) and confirmed by nuclear magnetic resonance studies (*37*).

The specific pattern of hydrophobic and hydrophilic residues in an amino acid sequence must limit the number of different structures a given sequence can adopt and may indeed define its overall fold. If this is true, then the arrangement of hydrophobic and hydrophilic residues should be a characteristic feature of a particular fold. Sweet and Eisenberg have shown that the correlation of the pattern of hydrophobicity between two protein sequences is a good criterion for their structural relatedness (*38*). In addition, several studies indicate that patterns of obligatory hydrophobic positions identified from aligned sequences are distinctive features of sequences that adopt the same structure (*4*, *29*, *38*, *39*). Thus, the order of hydrophobic and hydrophilic residues in a sequence may actually be sufficient information to determine the basic folding pattern of a protein sequence.

Although the pattern of sequence hydrophobicity may be a characteristic feature of a particular fold, it is not yet clear how such patterns could be used for prediction of structure de novo. It is important to understand how patterns in sequence space can be related to structures in conformation space. Lau and Dill have approached this problem by studying the properties of simple sequences composed only of H (hydrophobic) and P (polar) groups on two-dimensional lattices (*40*). An example of such a representation is shown in Fig. 5. Residues adjacent in the sequence must occupy adjacent squares on the lattice, and two residues cannot occupy the same space. Free energies of particular conformations are evaluated with a single term, an attraction of H groups. By considering chains of ten residues, an exhaustive conformational search for all 1024 possible sequences of H and P residues was possible. For longer sequences only a representative fraction of the allowed sequence or conformation space could be explored. The significant results were as follows: (i) not all sequences can fold into a "native" structure and only a few sequences form a unique native structure; (ii) the probability that a sequence will adopt a unique native structure increases with chain length; and (iii) the native states are compact, contain a hydrophobic core surrounded by polar residues, and contain significant secondary structure. Although the gap between these two-dimensional simulations and three-dimensional structures is large, the use of simple rules and sequence representations yields results similar to those expected for real proteins. Three-dimensional lattice methods are also beginning to be developed and evaluated (*41*).

Summary

There is more information in a set of related sequences than in a single sequence. A number of practical applications arise from an analysis of the tolerance of residue positions to change. First, such information permits the evaluation of a residue's importance to the function and stability of a protein. This ability to identify the essential elements of a protein sequence may improve our understanding of the determinants of protein folding and stability as well as protein function. Second, patterns of tolerance to amino acid substitutions of varying hydrophilicity can help to identify residues likely to be buried in a protein structure and those likely to occupy

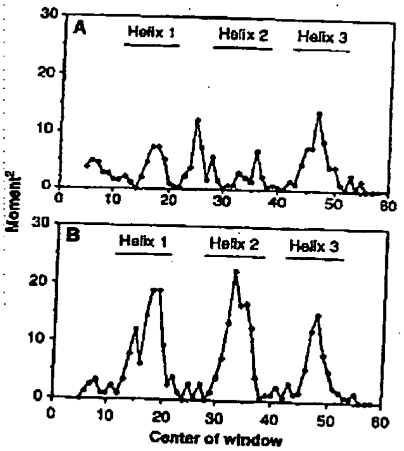

Fig. 4. Helical hydrophobic moments calculated by using (**A**) the Antennapedia homeodomain sequence or (**B**) a set of 39 aligned homeodomain sequences (*35*). The bars indicate the extent of the helical regions identified in nuclear magnetic resonance studies of the Antennapedia homeodomain (*36*). To determine hydrophobic moments, residues were assigned to one of three groups: H1 (high hydrophobicity = Trp, Ile, Phe, Leu, Met, Val, or Cys); H2 (medium hydrophobicity = Tyr, Pro, Ala, Thr, His, Gly, or Ser); and H3 (low hydrophobicity = Gln, Asn, Glu, Asp, Lys, or Arg). For the aligned homeodomain sequences, the residues at each position were sorted by their hydrophobicity by using the scale of Fauchere and Pliska (*45*). Arg and Lys were not counted unless no other residue was found at the position, because they contain long aliphatic side chains and can thereby substitute for nonpolar residues at some buried sites. To account for possible sequence errors and rare exceptions, the most hydrophilic residue allowed at each position was discarded unless it was observed twice. The second most hydrophilic residue was then chosen to represent the hydrophobicity of each position. An eight-residue window was used and the vectors projected radially every 100°. The vector magnitudes were assigned a value of 1, 0, or −1 for positions where the hydrophobicity group was H1, H2, or H3, respectively.

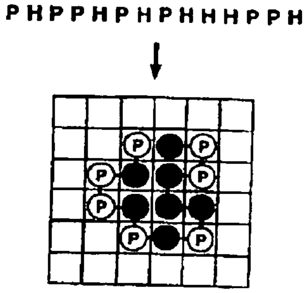

Fig. 5. A representation of one compact conformation for a particular sequence of H and P residues on a two-dimensional square lattice. [Adapted from (*40*), with permission of the American Chemical Society]

surface positions. The amphipathic patterns that emerge can be used to identify probable regions of secondary structure. Third, incorporating a knowledge of allowed substitutions can improve the ability to detect and align distantly related proteins because the essential residues can be given prominence in the alignment scoring.

As more sequences are determined, it becomes increasingly likely that a protein of interest is a member of a family of related sequences. If this is not the case, it is now possible to use genetic methods to generate lists of allowed amino acid substitutions. Consequently, at least in the short term, it may not be necessary to solve the folding problem for individual protein sequences. Instead, information from sequence sets could be used. Perhaps by simplifying sequence space through the identification of key residues, and by simplifying conformation space as in the lattice methods, it will be possible to develop algorithms to generate a limited number of trial structures. These trial structures could then, in turn, be evaluated by further experiments and more sophisticated energy calculations.

REFERENCES AND NOTES

1. C. J. Epstein, R. F. Goldberger, C. B. Anfinsen, *Cold Spring Harbor Symp. Quant. Biol.* **28**, 439 (1963); C. B. Anfinsen, *Science* **181**, 223 (1973).
2. R. E. Dickerson, *Sci. Am.* **242**, 136 (March 1980).
3. M. D. Hampsey, G. Das, F. Sherman, *FEBS Lett.* **231**, 275 (1988).
4. D. Bashford, C. Chothia, A. M. Lesk, *J. Mol. Biol.* **196**, 199 (1987).
5. A. M. Lesk and C. Chothia, *ibid.* **136**, 225 (1980).
6. M. F. Perutz, J. C. Kendrew, H. C. Watson, *ibid.* **13**, 669 (1965).
7. C. Chothia and A. M. Lesk, *Cold Spring Harbor Symp. Quant. Biol.* **52**, 399 (1965).
8. J. U. Bowie and R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2152 (1989).
9. J. F. Reidhaar-Olson and R. T. Sauer, *Science* **241**, 53 (1988); *Proteins Struct. Funct. Genet.*, in press.
10. D. Shortle, *J. Biol. Chem.* **264**, 5315 (1989).
11. J. H. Miller *et al.*, *J. Mol. Biol.* **131**, 191 (1979).
12. S. Sprang *et al.*, *Science* **237**, 905 (1987); C. S. Craik, S. Roczniak, C. Largman, W. J. Rutter, *ibid.*, p. 909.
13. H. C. M. Nelson and R. T. Sauer, *J. Mol. Biol.* **192**, 27 (1986).
14. M. H. Hecht, J. M. Sturtevant, R. T. Sauer, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 5685 (1984).
15. T. Alber, D. Sun, J. A. Nye, D. C. Muchmore, B. W. Matthews, *Biochemistry* **26**, 3754 (1987).
16. D. Shortle and A. K. Meeker, *Proteins Struct. Funct. Genet.* **1**, 81 (1986).
17. A. M. Lesk and C. Chothia, *J. Mol. Biol.* **160**, 325 (1982).
18. W. R. Taylor, *ibid.* **188**, 233 (1986).
19. W. Kauzmann, *Adv. Protein Chem.* **14**, 1 (1959); R. L. Baldwin, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 8069 (1986).
20. W. A. Lim and R. T. Sauer, *Nature* **339**, 31 (1989); in preparation.
21. Lesk and Chothia (*5*) have argued that a protein core composed solely of hydrogen-bonded residues would also be inviable on evolutionary grounds, as a mutational change in one core residue would require compensating changes in any interacting residue or residues to maintain a stable structure.
22. T. M. Gray and B. W. Matthews, *J. Mol. Biol.* **175**, 75 (1984); E. N. Baker and R. E. Hubbard, *Prog. Biophys. Mol. Biol.* **44**, 97 (1984).
23. F. M. Richards, *J. Mol. Biol.* **82**, 1 (1974).
24. J. W. Ponder and F. M. Richards, *ibid.* **193**, 775 (1987).
25. J. T. Kellis, Jr., K. Nyberg, A. R. Fersht, *Biochemistry* **28**, 4914 (1989); W. S. Sandberg and T. C. Terwilliger, *Science* **245**, 54 (1989).
26. A. A. Pakula and R. T. Sauer, *Proteins Struct. Funct. Genet.* **5**, 202 (1989).
27. B. C. Cunningham and J. A. Wells, *Science* **244**, 1081 (1989); R. M. Breyer and R. T. Sauer, *J. Biol. Chem.* **264**, 13348 (1989).
28. B. C. Cunningham, P. Jhurani, P. Ng, J. A. Wells, *Science* **243**, 1330 (1989).
29. L. H. Pearl and W. R. Taylor, *Nature* **329**, 351 (1987).
30. W. J. Brown *et al.*, *J. Mol. Biol.* **42**, 65 (1969); J. Greer, *ibid.* **153**, 1027 (1981); J. M. Berg, *Proc. Natl. Acad. Sci. U.S.A.* **85**, 99 (1988).
31. W. R. Taylor, *Protein Eng.* **2**, 77 (1988).
32. M. A. Navia *et al.*, *Nature* **337**, 615 (1989).
33. M. Schiffer and A. B. Edmundson, *Biophys. J.* **7**, 121 (1967); V. I. Lim, *J. Mol. Biol.* **88**; 857 (1974); *ibid.*, p. 873.
34. D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Nature* **299**, 371 (1982); D. Eisenberg, D. Schwarz, M. Komaromy, R. Wall, *J. Mol. Biol.* **179**, 125 (1984); D. Eisenberg, R. M. Weiss, T. C. Terwilliger, *Proc. Natl. Acad. Sci. U.S.A.* **81**, 140 (1984).
35. T. R. Burglin, *Cell* **53**, 339 (1988).
36. G. Otting *et al.*, *EMBO J.* **7**, 4305 (1988).
37. J. N. Breg, R. Boelens, A. V. E. George, R. Kaptein, *Biochemistry* **28**, 9826 (1989); M. G. Zagorski, J. U. Bowie, A. K. Vershon, R. T. Sauer, D. J. Patel, *ibid.*, p. 9813.
38. R. M. Sweet and D. Eisenberg, *J. Mol. Biol.* **171**, 479 (1983).
39. J. U. Bowie, N. D. Clarke, C. O. Pabo, R. T. Sauer, *Proteins Struct. Funct. Genet.*, in preparation.
40. K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
41. A. Sikorski and J. Skolnick, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 2668 (1989); A. Kolinski, J. Skolnick, R. Yaris, *Biopolymers* **26**, 937 (1987); D. G. Covell and R. L. Jernigan, *Biochemistry*, in press.
42. B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971).
43. S. R. Jordan and C. O. Pabo, *Science* **242**, 893 (1988).
44. R. M. Breyer, thesis, Massachusetts Institute of Technology, Cambridge (1988).
45. J.-L. Fauchere and V. Pliska, *Eur. J. Med. Chem.-Chim. Ther.* **18**, 369 (1983).

46. We thank C. O. Pabo and S. Jordan for coordinates of the NH_2-terminal domain of λ repressor and its operator complex. We also thank P. Schimmel for the use of his graphics system and J. Burnbaum and C. Francklyn for assistance. Supported in part by NIH grant AI-15706 and predoctoral grants from NSF (J.R.-O.) and Howard Hughes Medical Institute (W.A.L.).